Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2010 Proceedings

Americas Conference on Information Systems (AMCIS)

8-2010

Quality of XBRL US GAAP Taxonomy: Empirical Evaluation using SEC Filings

Hongwei Zhu
Old Dominion University, hzhu@odu.edu

Harris Wu
Old Dominion University, hwu@odu.edu

Quality of XBRL US GAAP Taxonomy: Empirical Evaluation using SEC Filings

Hongwei Zhu
College of Business and Public Administration
Old Dominion University
hzhu@odu.edu

Harris Wu
College of Business and Public Administration
Old Dominion University
hwu@odu.edu

ABSTRACT

The primary purpose of a data standard is to improve the comparability of data created by multiple standard users. Given the high cost of developing and implementing data standards, it is desirable to be able to assess the quality of data standards. We develop metrics for measuring completeness and relevancy of a data standard. These metrics are evaluated empirically using the US GAAP taxonomy in XBRL and SEC filings produced using the taxonomy by approximately 500 companies. The results show that the metrics are useful and effective. Our analysis also reveals quality issues of the GAAP taxonomy and provides useful feedback to the taxonomy users. The SEC has mandated that all publicly listed companies must submit their filings using XBRL beginning mid 2009 to late 2014 according to a phased-in schedule. Thus our findings are timely and have practical implications that will ultimately help improve the quality of financial data.

Keywords (Required)

Data quality, interoperability, data standards, XBRL, GAAP, long tail.

INTRODUCTION

From day-to-day operations to strategic decision making, organizations need data from disparate sources both internally and externally. They have to reconcile the heterogeneity in terms of format, semantics, and quality of data from these sources. This has been a challenging task for organizations and a difficult problem for the data integration research community for years. Many of the problems can be avoided if a data standard is used by all sources. Indeed, data standardization has the potential of ensuring quality and enhancing interoperability of data from disparate sources. There have been successful large-scale data standardization efforts such as those within the Department of Defense (DoD) (Rosenthal et al. 2004) and across the real estate mortgage industry (Markus et al. 2006). Most data collection efforts organized by the government also rely on data standards.

Does a data standard always improve the quality, especially the interoperability, of data created by different organizations using the standard? This is an important question given that it is often very costly to develop and implement a data standard. In this paper, we investigate this question empirically by analyzing the financial statements of nearly 500 companies produced in eXtensible Business Markup Language (XBRL) (XBRL International 2006) using the US GAAP taxonomy, which is a data standard also encoded using XBRL.

The GAAP taxonomy published in January 2009 defines more than 13,000 data elements that companies can use to produce their financial statements and file them to the Securities and Exchange Commission (SEC). One of the purposes of the standard GAAP taxonomy is to promote interoperability amongst financial statements of different companies. There are several aspects of data interoperability. From a syntactical perspective, all XBRL documents prepared using GAAP taxonomy and its extensions are interoperable because they use the same syntax and data typing system. But from a semantic perspective, XBRL documents from different companies can be difficult to compare when different companies use different data elements in their documents. This is because the semantic correspondence between elements in different documents is difficult to establish either computationally or manually. SEC allows a company to extend the standard taxonomy by defining its own elements in financial statements. Thus whether the GAAP taxonomy has helped increase the comparability of financial statements across companies depends on how companies use and extend the taxonomy. In this paper, we will focus on the semantic aspect of interoperability. In addition, we will use the term data and information interchangeably.

Data quality is defined as data's fitness for use (Wang et al. 1996). A data standard is meta-data that specifies the characteristics of data elements and their relationships. From the perspective of standard users, meta-data is also data. Thus data quality concepts also apply to data standards. Since one of the primary purposes of data standards is to produce highly

interoperable data, quality of data standards can be assessed by the interoperability of the resulting data. We will extend the notion of standard quality introduced in (Zhu et al. 2009) and evaluate it using real-world XBRL filings.

BCKGROUND

In this section, we provide a brief introduction to XBRL and the concept of data standard quality.

XBRL

XBRL as a language is defined in XBRL Specification (XBRL International 2006). It offers syntactic uniformity (e.g., a standard set of data types) desired for automatic data processing. The language can be used to develop XBRL taxonomies in different jurisdictions. These taxonomies are essentially different XBRL data standards. In the U.S., the standard taxonomy adopted by the SEC is the U.S. GAAP taxonomy.

XBRL taxonomies define a set of concepts (e.g., *operating profit*) as XML elements. For each element, its data type, attributes, relationships with other elements, and relationships with other resources (namely labels for human readers and references to authoritative sources). Element specification is provided using XML Schema, which specifies element name, data types, and other XBRL-specific attributes. Below is an example specification in the GAAP taxonomy for the element StockholdersEnquity:

```
<xs:element id='us-gaap_StockholdersEquity' name='StockholdersEquity' nillable='true'
substitutionGroup='xbrli:item' type='xbrli:monetaryItemType' xbrli:balance='credit'
xbrli:periodType='instant' />
```

Element relationships are specified using five XLINK-based *linkbases*:

- A definition linkbase specifies the conceptual prelateships between elements, mainly the generalization-specialization relationship often found in OO, extended ER, and ontology modeling.
- A calculation linkbase defines the numeric relationships between elements.
- A presentation linkbase specifies the hierarchical grouping (mainly the parent-child relationship) and the order of the elements when they are presented in a report for viewing purposes.
- A label linkbase provides the human-readable documentation for the elements defined in the taxonomy schema.
- A reference linkbase provides further explanations to the elements by linking them to authoritative references (e.g., SEC regulations or certain accounting standards) that define the meaning of the elements.

A company can extend a standard taxonomy in various ways, such as adding new data types, overriding element specifications, and adding custom elements. The most commonly practiced extension is to add custom elements.

A company's financial statement is an XML document that contains the facts tagged using the elements defined in taxonomies (including standard and company extension). In addition, each fact is associated with a context, which specifies the entity related to the fact and the time period of the fact. If the fact is of a numeric type, it is also associated with a unit of measure. Below is an example fact in a company filing to the SEC showing that the StockholdersEquity is $43.641B:

```
<us-gaap:StockholdersEquity contextRef="eol_PE9932----0910-
K0004_STD_0_20061231_0_411810x401098" unitRef="iso4217_USD" decimals="-6">43641000000
</us-gaap:StockholdersEquity>
```

Quality of Data Standards

Most data quality research focuses on data, not the standards used to create and organize the data. Data quality is a multi-dimensional concept that goes beyond accuracy. Prior research has identified 16 dimensions (e.g., consistency, interpretability, completeness, relevancy, etc.) of data quality (Wang et al. 1996). Data quality perceived by users of different roles within an organization can be assessed using survey instruments (Lee et al. 2002b). Quality of database schemas is discussed in (Redman 1996). Although a database schema is a type of data standard, it is mainly used within a single organization to organize and store data in a database. In contrast, the main objective of many data standards is to allow for meaningful exchange of data among multiple organizations so that the data from different organizations are interoperable.

The notion of data standard quality and two metrics that measure standard relevancy and completeness are introduced in (Zhu et al. 2009). Below we extend that work by providing definitions for the notion and metrics.

The quality of a data standard is its fitness for multiple users to produce highly interoperable data. Like data quality, data standard quality has multiple dimensions. Further research is needed to determine these dimensions. At the minimum, it has completeness and relevancy dimensions.

For data quality, completeness is defined as "the extent to which data are of sufficient breadth, depth, and scope for the task at hand", and relevancy is defined as "the extent to which data are applicable and helpful for the task at hand" (Wang et al. 1996). Schema completeness and pertinence (i.e., relevancy) are defined similarly in (Redman 1996). These definitions need to be adapted for data standard quality. Our definitions are:

- Completeness of a data standard is the extent to which the data standard specifies all the data elements needed by standard users.
- Relevancy of a data standard is the extent to which the data standard specifies only the data elements needed by standard users.

The completeness and relevancy of the same data standard can be different to different users. Further, they can be different between an individual user and the user community. To formalize the metrics, let the **S** be the set of data elements specified in the data standard, $\mathbf{U_i}$ be the data elements required by the user i. From the user i's perspective, the metrics can be defined as

$$Completeness_i = \frac{|U_i \cap S|}{|U_i|}, \text{ and } Relevancy_i = \frac{|U_i \cap S|}{|S|}$$

From the user community's perspective, the metrics can be defined as

$$Completeness_c = \frac{\left|\left(\bigcup_i U_i\right) \cap S\right|}{\left|\bigcup_i U_i\right|}, \text{ and } Relevancy_c = \frac{\left|\left(\bigcup_i U_i\right) \cap S\right|}{|S|}$$

A standard can be complete by specifying every possible data elements, but it will suffer from low relevancy because many of the specified data elements may not be needed by any user. Conversely, a standard can be highly relevant by only specifying crucial data elements that are absolutely needed by all users, but it may be incomplete because it does not specify certain data elements occasionally needed by a few users.

A measure that combines completeness and relevancy is the harmonic mean of the completeness and relevancy:

F = 2*completeness*relevance / (completeness + relevancy)

The above measure is analogical to the classic F-measure, often used to evaluate the effectiveness of information retrieval. Completeness and relevancy are called "recall" and "precision", respectively, in the information retrieval literature (van Rijsbergen 1979).

For different contexts users would value completeness and relevancy differently. For example, when it is extremely difficult for non-standard data to interoperate, the completeness of the standard will be more useful. For example, a typical language in any culture would contain tens of thousands of words for completeness. However if human cognitive load is a critical factor in adoption and effective usage of standards, relevancy is more useful. For example, maritime sign language only consists of dozens of basic signs.

To accommodate different weights users place on the importance of completeness and relevancy, we define the quality of a standard for a user (or group of users) as a general F_β measure (for non-negative real values of β):

$$F_\beta = (1+\beta^2)\frac{relevancy \bullet completeness}{\beta \bullet relevancy + completeness}$$

For example, F_2 measure weights completeness twice as much as relevancy, and the $F_{0.5}$ measure weights relevancy twice as much as completeness. The F-measure was derived from van Rijsbergen (1979) where F_β "measures the effectiveness of retrieval with respect to a user who attaches β times as much importance to recall as precision".

RESEARCH METHOD

We evaluate the concepts of data standard quality empirically using the GAAP taxonomy and XBRL filings submitted to the SEC. The SEC mandated that all publically listed companies must submit their filings in XBRL by October 31, 2014. The mandate takes places in phases, beginning with large companies who must use XBRL as of June 15, 2009. Our dataset contains all required SEC filings from June 15, 2009 until February 26, 2010. It contains 1,231 filings submitted by 478 companies, among which 463 have more than one filing (e.g., multiple 10-Q and 10-K filings). All of the companies used the GAAP Taxonomy version 2009-01-31.

In addition to evaluating the concepts of data standard quality, this empirical study also has timely practical value to the SEC, accounting professionals, and financial analysts who produce and consume XBRL data. More importantly, it offers insights for the GAAP taxonomy community to improve the quality of the data standard.

Our data acquisition, processing, and analysis methods are depicted in Figure 1.

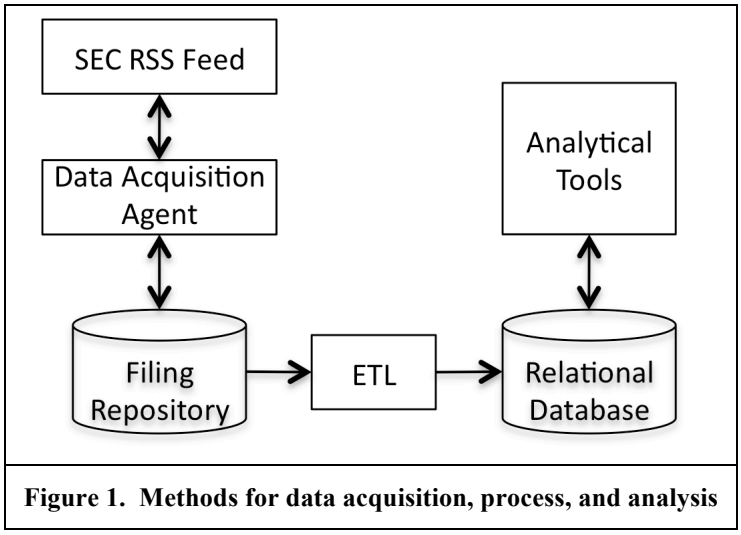

Figure 1. Methods for data acquisition, process, and analysis

We have developed a data acquisition agent that monitors the RSS Feed at SEC and other sites to obtain company filings submitted to the SEC. The acquisition agent downloads the financial statements and the accompanying taxonomy extensions into a local filing repository. The ETL program parses the files downloaded and loads the extracted data into a relational database. Stored SQL procedures and other programs are used to analyze the data stored in the relational database.

The quality metrics defined earlier use the set notion. In XML and XBRL, a data element is identified by its name and name space. When a company extends the standard taxonomy by introducing new data elements, the elements have a name space unique to the company. Thus even two companies use the same name for their data elements, the elements are different because they have different name spaces. However, when two elements names are the same or similar, it is highly likely that they are semantically equivalent. As an initial step to identify such potential semantic equivalents from the tens of thousands of data elements used by the companies, we use the cosine similarity of element names that are represented by word vectors. In the element vectors, words are weighted according to the traditional term-frequency / inverse document frequency (TF/IDF) weighting in the computational linguistics literature. A word is a term, whereas the name of an element is a "document". Words that appear in many different elements, such as "Net" and "Stock", are given a low weight. Words that appear multiple times in a given element name are given multiple weights.

EMPIRICAL FINDINGS

GAAP TAXONOMY

The GAAP taxonomy specifies a total of 13,452 data elements, among which 2,653 are abstract and 346 are deprecated. Abstract elements are used for deriving concrete data elements and cannot be used in company filings. All depreciated elements are deprecated on January 31, 2009, meaning that these elements are not recommended to use in filings after January 31, 2009. Among the abstract elements, 84 are deprecated. Thus the number of concrete elements is 10,799, of which 10,537 are active (i.e., not depreciated).

Using our similarity tool we found that many of the deprecated elements have a corresponding active element. Their names tend to be permutations of the same set of words. For example, the deprecated element CashDividends has a corresponding active element DividendsCash, and the depreciated element CashProvidedByUsedInDiscontinuedOperations*FinancingActivities* has a corresponding active element CashProvidedByUsedIn*FinancingActivities*DiscontinuedOperations. As highlighted using underline and italic font the second two elements are different in the sequence of DiscontinuedOperations and FinancingActivities. In total there are 5 cases where deprecated elements are word-permutations of active elements.

Many companies continued to use deprecated standard elements. Overall 40.5% of the companies (195 out of 481) used deprecated elements in 19.82% of the filings (244 out of 1231). Below are a list of deprecated elements and the number of filings that still used them.

AccountsPayable	137
CommitmentsAndContingencies	121
OtherAdjustmentsForNoncashItemsIncludedInIncomeLossFromContinuingOperations	89
AccruedLiabilities	72
AccruedIncomeTaxesPayable	60
EmployeeRelatedLiabilities	48
CashDividends	43
AccountsPayableAndAccruedLiabilities	42
PensionAndOtherPostretirementDefinedBenefitPlansNoncurrentLiabilities	35
TaxesPayable	28
MinorityInterestInNetIncomeLossOfConsolidatedEntities	26
InterestPayable	24
TaxesOtherThanIncomeExciseProductionAndPropertyTaxes	23
OtherPostretirementDefinedBenefitPlanNoncurrentLiabilities	21
DefinedBenefitPensionPlanNoncurrentLiabilities	19
OtherAccruedLiabilities	19
DividendsPayable	13
CashProvidedByUsedInDiscontinuedOperationsOperatingActivities	8
StockConvertedFromOneClassToAnotherClassValue	8
DueToRelatedParties	8
ScheduleOfAccountsAndNotesReceivableTextBlock	8
CommonStockAdditionalSeriesValue	6
AccountsPayableTrade	6
CommonStockAdditionalSeriesSharesIssued	6
WriteOffOfInventory	6
CommonStockAdditionalSeriesSharesAuthorized	5
CommonStockAdditionalSeriesParOrStatedValuePerShare	5
AccruedRoyalties	5
CashProvidedByUsedInDiscontinuedOperationsInvestingActivities	5
ConvertiblePreferredStockSharesIssued	5
ScheduleOfCommonStockByClassTextBlock	5
ConvertiblePreferredStockParStatedValuePerShare	5
AdjustmentsToReconcileIncomeLossToNetCashProvidedByUsedInContinuingOperations	5
CashProvidedByUsedInDiscontinuedOperationsFinancingActivities	4
InterestAndDividendsPayable	4
DeferredCompensationLiabilityNoncurrent	4
ConvertiblePreferredStockSharesOutstanding	4
RelatedPartyDebtNoncurrent	4
ConvertiblePreferredStockSharesAuthorized	4
RevenueFromLeaseOrRentalOfPropertyOrEquipment	4
ProfessionalAndContractServicesExpense	4
DueToAffiliate	3
CommercialPaperCurrent	3
StockConvertedFromOneClassToAnotherClassShares	3
AccrualForTaxesOtherThanIncomeTaxes	3

ProductWarrantyAccrualCurrent	3
DefinedBenefitPlanNoncurrentAssetsForPlanBenefits	3
AmortizationOfDeferredAcquisitionCostsDAC	3
CommonStockAdditionalSeriesSharesOutstanding	3
PreferredStockAdditionalSeriesValue	2
PensionAndOtherPostretirementAndPostemploymentBenefitPlansNoncurrentLiabilities	2
InventoryPartsAndComponents	2
NotesPayableRelatedPartiesCurrent	2
ScheduleOfPreferredStockByClassTextBlock	2
AccountsPayableRelatedParties	2
AntidilutiveSharesOutstanding	2
LongTermDebtComponentsMortgageLoans	2
StockIssuedDuringPeriodValueSharesHeldInTrustOfEmployeeStockOwnershipPlan	2
AccruedSalaries	2
AdjustmentsToReconcileToIncomeLossFromContinuingOperations	1
PreferredStockAdditionalSeriesSharesOutstanding	1
PreferredStockAdditionalSeriesSharesIssued	1
AccruedAdvertising	1
PreferredStockAdditionalSeriesSharesAuthorized	1
SalesAndExciseTaxPayable	1
DepreciationExpense	1
ProductionAndPropertyTaxExpense	1
AccruedInsurance	1
AdjustmentsForNoncashItemsIncludedInIncomeLossFromContinuingOperations	1
PaymentsForMergerRelatedCostsAndRestructuringCosts	1
ProceedsFromSaleOfLand	1
AccountsPayableOther	1
AccruedSalesCommission	1
PreferredStockAdditionalSeriesLiquidationPreference	1

Company Filings

Standard Quality: Completeness and Relevancy

Of the 1,231 filings in the dataset, 266 are 10-K (annual), 964 are 10-Q (quarterly), and 1 is 20-F (filed by an Israeli pharmaceutical company). All the filings use the GAAP taxonomy as the base taxonomy. For each filing, we identify data elements specified in the GAAP taxonomy and those introduced by the filing company. The statistics of the numbers of two types of elements is given in Table 1.

	Taxonomy	Custom	Both
Min	17	0	20
Max	246	7	385
Mean	109.5	15.6	125.1
Median	106	12	120
Stand deviation	24.1	13.7	32.3

Table 1. Statistics of number of elements per filing

Generally, more taxonomy elements were used than custom elements in each filing. On average, a company filing used 109.5 elements from the GAAP taxonomy and 15.6 custom elements.

All the companies together used 2,558 GAAP elements and introduced 10,168 custom elements. Using the metrics defined earlier, we can compute the completeness and relevancy of the GAAP taxonomy from the perspective of the average filing company and from the perspective of all filing companies (i.e., the user community). As discussed earlier, certain companies used deprecated elements. Even though all filings were submitted after the date of the deprecation date, it is not compulsory to refrain from using them. Thus we use all concrete elements for |S|, which is 10,799. The results are shown in Table 2.

	Completeness	Relevancy	F-measure
Average company	109.5/125.1=0.8753	109.5/10799=0.0101	0.02
User community	2585/(2558+10168)=0.2027	2585/10799=0.2394	0.22

Table 2. Completeness and relevancy of GAAP taxonomy from perspectives of average user and user community

For the average company, the GAAP taxonomy has a high completeness score because most elements in a filing are from the GAAP taxonomy. But the relevancy score is extremely low because the number of elements used by the average company only represents 1% of the elements specified in the taxonomy. Identifying these desired elements can be a daunting task. For the user community, the completeness and relevancy scores are approximately 0.2. These values will change as more companies use the taxonomy. For example, when companies start to use more elements in the taxonomy, the relevancy will increase. It is uncertain whether the completeness score will increase or decrease in the future, because it depends on the number of custom elements that companies will introduce. In our next step of research, we will investigate whether the standard quality metrics have changed between the voluntary filings before March 2009 and the official filings since June 2009. The companies' adoption of and familiarity with the GAAP standard certainly has an impact on the perceived standard quality. Further examination of the usage of the GAAP and custom elements could also suggest directions for quality improvement. For example, incorporating certain custom elements into the GAAP taxonomy may improve the quality metrics. In the section below, we examine the element usage, similarity and complexity.

Element Usage Frequency

Now let us look at the use frequencies of the taxonomy elements and custom elements. An element can appear in a company filing more than once. This is because a filing often contains data for multiple reporting periods and certain data elements appear once for each time period. For the purpose of counting use frequency of data elements in filings, we use binary counting method were if an element occurs in a filing the count is incremented by 1 regardless of the number of occurrences of the element in the filing. For company-introduced elements, we treat the same-named elements used by different companies as the same element, disregarding the name space. This counting method essentially assumes the elements with the same name are the same elements. The use frequencies of the taxonomy elements and custom elements are illustrated in Figure 2.

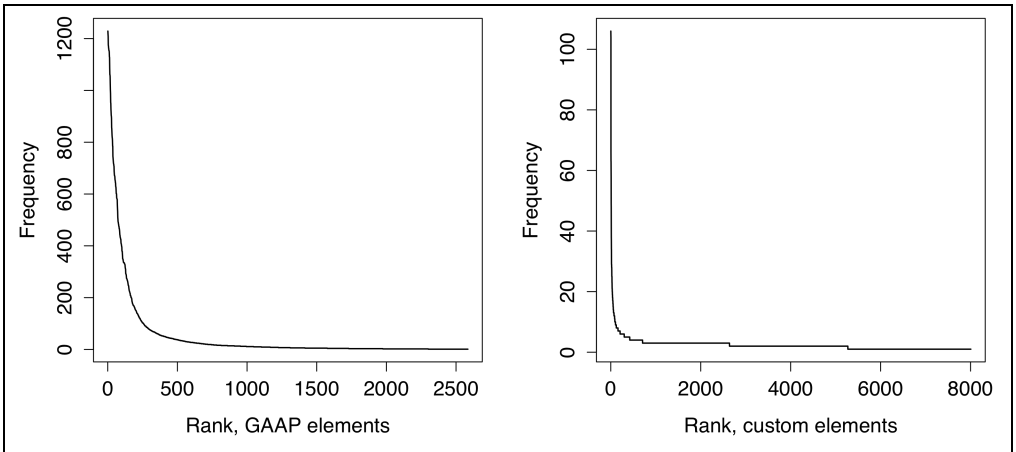

Figure 2. Use frequencies of GAAP taxonomy elements and custom elements

In the figure, the Y-axis is the use frequency, the X-axis of the frequency rank. It appears that both types of the elements have a power law distribution, also known as the long-tail distribution. There are a few elements that are used frequently, but most of the elements are only used in a small number of company filings. The top 50 most used elements from the GAAP taxonomy and introduced by companies are listed in Tables 3 and 4, respectively.

GAAP Element	Frequency
Assets	1229
LiabilitiesAndStockholdersEquity	1217
CashAndCashEquivalentsPeriodIncreaseDecrease	1209
IncomeTaxExpenseBenefit	1175

CashAndCashEquivalentsAtCarryingValue	1172
NetCashProvidedByUsedInInvestingActivities	1164
RetainedEarningsAccumulatedDeficit	1161
NetCashProvidedByUsedInFinancingActivities	1155
NetCashProvidedByUsedInOperatingActivities	1154
EarningsPerShareBasic	1153
EarningsPerShareDiluted	1147
PropertyPlantAndEquipmentNet	1130
AccumulatedOtherComprehensiveIncomeLossNetOfTax	1127
CommonStockValue	1086
NetIncomeLoss	1061
AssetsCurrent	1060
LiabilitiesCurrent	1053
StockholdersEquity	1024
OtherAssetsNoncurrent	1002
SegmentReportingDisclosureTextBlock	989
Goodwill	954
OtherLiabilitiesNoncurrent	950
CommonStockParOrStatedValuePerShare	927
CommonStockSharesAuthorized	917
CommitmentsAndContingenciesDisclosureTextBlock	904
OperatingIncomeLoss	901
CommonStockSharesIssued	893
WeightedAverageNumberOfDilutedSharesOutstanding	857
WeightedAverageNumberOfSharesOutstandingBasic	852
IncomeTaxDisclosureTextBlock	839
IncomeLossFromContinuingOperationsBeforeIncomeTaxesMinorityInterestAndIncomeLossFromEquityMethodInvestments	826
PaymentsForRepurchaseOfCommonStock	819
ShareBasedCompensation	815
PaymentsToAcquirePropertyPlantAndEquipment	807
TreasuryStockValue	785
InterestExpense	785
InventoryNet	746
DeferredIncomeTaxExpenseBenefit	741
Liabilities	734
PensionAndOtherPostretirementBenefitsDisclosureTextBlock	727
DerivativeInstrumentsAndHedgingActivitiesDisclosureTextBlock	720
EffectOfExchangeRateOnCashAndCashEquivalents	716
PaymentsForProceedsFromOtherInvestingActivities	715
FairValueDisclosuresTextBlock	709
EarningsPerShareTextBlock	707
StockholdersEquityIncludingPortionAttributableToNoncontrollingInterest	688
AccountsPayableCurrent	680
AccountsReceivableNetCurrent	672
Revenues	671
OtherAssetsCurrent	670

Table 3. Top 50 most used data elements in GAAP taxonomy

Custom Element	Frequency
IncomeLossFromContinuingOperationsBeforeIncomeTaxes	106
EarningsPerShareTextBlock	105
PrepaidExpensesAndOtherCurrentAssets	104
IncomeBeforeIncomeTaxes	66
TotalOtherAssets	63
StockholdersEquityIncludingPortionAttributableToNoncontrollingInterest	63
ProfitLoss	60
NetIncomeLossAttributableToNoncontrollingInterest	58
SharesOutstanding	49
InterestExpenseNet	48
TotalDeferredCreditsAndOtherLiabilities	42
PrepaidExpenseAndOtherAssetsCurrent	38
IncreaseDecreaseOtherCurrentAssets	37
ContributionsFromNoncontrollingInterests	34

BusinessDescriptionAndSignificantAccountingPoliciesTextBlock	34
DetailsOfCertainBalanceSheetAccountsDisclosureTextBlock	29
PrepaidExpensesAndOther	29
ContingenciesDisclosureTextBlock	29
NetOtherThanTemporaryImpairments	28
DescriptionOfNewAccountingPronouncementsNotYetAdoptedTextBlock	28
IncomeLossFromContinuingOperationsIncludingPortionAttributableToNoncontrollingInterest	28
IncreaseDecreaseInIncomeTaxesNet	27
AcquisitionsAndDispositionsDisclosureTextBlock	25
IncomeLossFromContinuingOperationsBeforeIncomeTaxesAndMinorityInterest	25
StockholdersEquitySubtotalBeforeTreasuryStock	25
BasisOfPresentationDisclosureTextBlock	25
EarningsPerShareBasicAndDiluted	23
ComprehensiveIncomeNetOfTaxIncludingPortionAttributableToNoncontrollingInterest	23
InvestmentsTextBlock	22
OtherComprehensiveIncomeOtherNetOfTax	21
IncreaseDecreaseOtherCurrentLiabilities	21
BasisOfPresentationTextBlock	21
OtherOperatingIncomeExpense	20
LegalProceedingsTextBlock	20
InvestmentsAndOtherAssets	20
AdjustmentDepreciationAndAmortization	19
IncomeFromContinuingOperationsBeforeIncomeTaxes	19
AdjustmentsToAdditionalPaidInCapitalSharebasedCompensationRequisiteServicePeriodRecognitionValue	19
OtherIncomeExpenseNet	18
AccruedExpensesAndOtherCurrentLiabilities	18
TotalCostsAndExpenses	18
NonCreditPortionOfOtherThanTemporaryImpairmentsRecognizedInOtherComprehensiveIncome	18
FinancialInstrumentsTextBlock	17
OtherThanTemporaryImpairments	17
MaterialsAndSupplies	17
InterestIncome	17
AdditionalFinancialInformationDisclosureTextBlock	17
OtherNet	16
ProceedsFromRepaymentsofCommercialPaper	16
IncomeLossBeforeIncomeTaxes	16

Table 4. Top 50 most used data elements introduced by filing companies

Company financial statements usually contain a set of common financial terms, thus one would expect that the XBRL filings contain many common data elements which should have be specified in GAAP taxonomy. Surprisingly, only the top five elements listed in Table 3 have been used in more than 95% of the filings and the 50th element has been used in only 54.4% of the filings.

Out of the top 50 custom data elements, 15 elements (or 30%) have the same names as those in the GAAP taxonomy. This could easily lead to confusion. We will investigate whether companies indeed have extended the original element from the GAAP taxonomy. But we suspect that in some cases companies likely have created duplicate elements.

The following list shows the frequency of an element name being used in a company namespace (first column) and in the GAAP taxonomy namespace (second column):

EarningsPerShareTextBlock	105	712
StockholdersEquityIncludingPortionAttributableToNoncontrollingInterest	63	689
ProfitLoss	60	655
NetIncomeLossAttributableToNoncontrollingInterest	58	490
ComprehensiveIncomeNetOfTaxIncludingPortionAttributableToNoncontrollingInterest	23	335
IncomeLossFromContinuingOperationsIncludingPortionAttributableToNoncontrollingInterest	28	249
AdjustmentsToAdditionalPaidInCapitalSharebasedCompensationRequisiteServicePeriodRecognitionValue	19	216
OtherComprehensiveIncomeOtherNetOfTax	21	3

DescriptionOfNewAccountingPronouncementsNotYetAdoptedTextBlock	28	3
ContributionsFromNoncontrollingInterests	34	3
EarningsPerShareBasicAndDiluted	23	1
PrepaidExpensesAndOther	29	1
BasisOfPresentationTextBlock	21	1
InterestExpenseNet	48	1
InterestIncome	17	1

Element Similarity

In addition to identical elements, many custom elements are very similar to the elements in the GAAP standard. Of the top 50 company-defined elements, 13 of them have a similar element in the GAAP taxonomy with a cosine similarity score greater than 0.8. The 13 custom elements (first column), the similar element in the GAAP taxonomy, and their similarity score are listed below:

PrepaidExpenseAndOtherAssetsCurrent	PrepaidExpenseCurrent	0.92
InterestExpenseNet	InterestIncomeExpenseNet	0.92
InterestExpenseNet	InterestExpenseOther	0.89
IncomeLossFromContinuingOperationsBeforeIncomeTaxesAndMinorityInterest	IncomeLossFromContinuingOperationsBeforeIncomeTaxesMinorityInterestAndIncomeLossFromEquityMethodInvestments	0.87
ComprehensiveIncomeNetOfTaxIncludingPortionAttributableToNoncontrollingInterest	OtherComprehensiveIncomeLossNetOfTaxPortionAttributableToNoncontrollingInterest	0.87
EarningsPerShareBasicAndDiluted	EarningsPerShareDiluted	0.85
AccruedExpensesAndOtherCurrentLiabilities	OtherAccruedLiabilitiesCurrent	0.83
AdjustmentDepreciationAndAmortization	DepreciationAndAmortization	0.83
ProceedsFromRepaymentsOfCommercialPaper	RepaymentsOfCommercialPaper	0.81
AccruedExpensesAndOtherCurrentLiabilities	AccruedLiabilitiesCurrent	0.81
IncomeFromContinuingOperationsBeforeIncomeTaxes	IncomeLossFromContinuingOperationsBeforeIncomeTaxesDomestic	0.81
AdjustmentDepreciationAndAmortization	OtherDepreciationAndAmortization	0.81
IncreaseDecreaseInIncomeTaxesNet	IncreaseDecreaseInIncomeTaxesReceivable	0.80

We notice the similarity score is not perfect. Further research will develop other methods to more accurately identify duplicates. Many more elements are similar to each other. The filings used 8006 company-defined elements and 2585 GAAP standard elements. Overall the filings have 10543 distinct named elements, and therefore 10543*10542/2 = 55,572,153 element pairs. We computed the cosine similarity among all these element pairs. The table below shows the similarity distribution.

Range	# of element pairs
[0,0.1)	53189258
[0.1,0.2)	1500122
[0.2,0.3)	517414
[0.3,0.4)	199096
[0.4,0.5)	86348
[0.5,0.6)	40902
[0.6,0.7)	20500
[0.7,0.8)	10404
[0.8,0.9)	5311
[0.9,1)	2711
1	87

Element Names and Complexity

Although element names are mainly used by computer tools, they are also used manually by humans when computer tools cannot accurately identify the desired elements. Thus it is useful to examine the syntactic and linguistic characteristics of element name. For the initial step, we use number of characters and number of words of element names. Figure 3 shows the histograms of these two metrics for both the taxonomy elements and custom elements. It appears that custom elements and GAAP elements have similar character and word length distributions. Also note that all these distributions are skewed long-tail distributions. Certain elements are very long – close to 200 characters, up to 55 words.

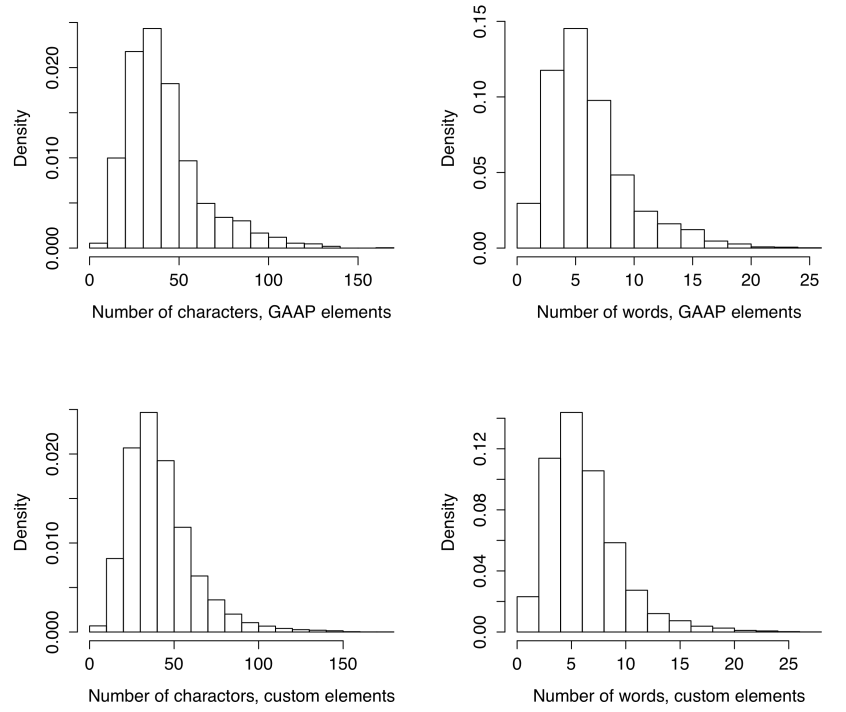

Figure 3. Element name length distributions

When drawing the histogram for the number of words of custom elements, we omitted two extreme cases: one with 38 words and the other with 55 words. The GAAP taxonomy also contained long element names up to 185 characters and 29 words.

Lengthy element names, although descriptive, require significant cogitative load to use them. In addition to large number of elements, lengthy names also contribute to the complexity of the standard taxonomy. We hypothesize that such complexity may induce unnecessary, duplicate elements to be introduced by taxonomy users. This hypothesis will be investigated in future research. The data shows that the lengthier-named elements are certainly less frequently used. For example, out of 39 elements in the GAAP taxonomy that contain 24 words or longer, only 1 element has ever been used in company filings. Figure 4 shows how the percentage of GAAP elements being used in the filings according to the length of the element name measured by the number of words and the number of characters. The general trend is that most of the short elements have been used in company filings. As the length increases, the percentage of the elements that have been used decreases.

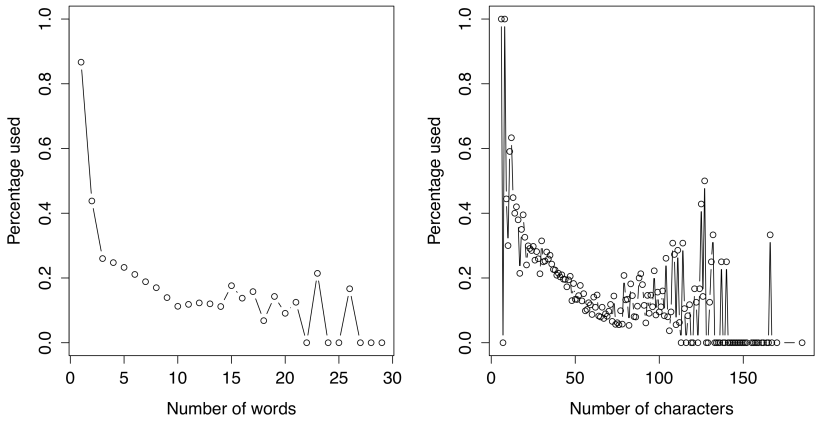

Figure 4. The percentage of GAAP elements being used decreases with the length of the element name

CONCLUSION AND FUTURE RESEARCH

As more data is produced and collected in digital format, it becomes more important that the data has high quality and interoperability. Data standards are often used in hopes of ensuring data quality. Thus it is critical to examine the impact of data standards on the quality of data. To this end, we must understand the concept of data standard quality and develop metrics and methods to measure the quality of data standards.

Building on extensive work on data quality, we have developed two separate metrics, completeness and relevancy, for data standard quality. The two metrics can be combined to produce the F-measure, where adjusted weights can be assigned to completeness and relevancy in different contexts. The metrics are applied to a real-world data standard. The results show that the metrics provide a useful measurement of the quality of data standards. Furthermore, our analysis of XBRL GAAP taxonomy and XBRL data has important and timely implications. It has been suggested that data standards such as the GAAP taxonomy should be evaluated using large datasets and automatic methods (Bovee et al. 2005). Prior research (Boritz et al. 2008a; Boritz et al. 2008b; Chou 2006; Debreceny et al. 2005; Zhu et al. 2009) analyzed XBRL data only in the SEC voluntary filing program that was in place prior to the SEC mandate of using XBRL. As more companies file their financial statements using XBRL, comparability of these statements are increasingly important for investors and the regulators to make decisions and enforce financial disclosure practices. A data standard with poor quality can lead to low comparability of financial statements, destroying the very purpose of creating the standard in the first place. Our preliminary findings seem to suggest that certain quality problems may be avoided by controlling the complexity of the data standard.

Despite many exciting findings reported here, this research is only at a very early stage. There are many areas where we plan to improve in future research. The similarity analysis used here is very simple. More sophisticated algorithms developed for schema matching (Rahm et al. 2001; Rahm et al. 2004) can be adapted to detect potential duplicate elements more accurately. These algorithms need to be extended to take advantages of rich heuristics in the "linkbases", part of the XBRL taxonomy that defines relationships among the elements.

The GAAP taxonomy is fairly new to filing companies. Through learning, the companies may begin to use more elements from the taxonomy. Thus we should continue to gather and analyze new filings as they come in to see how companies learn and how the measurements of the metrics evolve over time. Further, company-introduced elements may converge over time. If this trend emerges, the GAAP taxonomy may consider adopting these converging elements introduced by the companies.

Survey method has been proven effective in gauging perceived quality of data (Lee et al. 2002a). We plan to conduct a survey on the quality of GAAP taxonomy. This will allow us to obtain additional information not observable in the company filings. The survey findings will also help us improve the methodology for developing data standards. Manual inspection of company filings can be useful for identifying specific deficiencies of the taxonomy and certain misuses of the taxonomy (Bovee et al. 2002). Thus we also plan to conduct case studies to examine company filings and their filing practice in more detail, such as how decisions are made on whether adopting standard elements or introducing custom elements. Companies of different industries and sizes may have different reporting needs. Therefore further analyses need to be done to examine the

filings by different industries and companies of different sizes. Such analyses will become feasible as more companies start to file their financials using XBRL.

In summary, we think we have opened an exciting area of information quality research. With our initial work we have made an important step towards developing methods for assessing data standard quality.

REFERENCES

1. Boritz, E.J., and No, W.G. "Auditing an XBRL Instance Document: The Case of United Technologies Corporation," University of Waterloo, 2008a.
2. Boritz, E.J., and No, W.G. "SEC's XBRL Voluntary Program on Edgar: The Case for Quality Assurance " in: *SSRN: http://ssrn.com/abstract=1163254*, 2008b.
3. Bovee, M., Ettredge, M.L., Srivastava, R.P., and Vasarhelyi, M.A. "Does the Year 2000 XBRL Taxonomy Accommodate Current Business Financial-Reporting Practice?," *Journal of Information Systems* (16:2) 2002, pp 165-182.
4. Bovee, M., Kogan, A., Nelson, K., Srivastava, R.P., and Vasarhelyi, M.A. "Financial Reporting and Auditing Agent with Net Knowledge (FRAANK) and eXtensible Business Reporting Language (XBRL)," *Journal of Information Systems* (19:1) 2005, pp 19-41.
5. Chou, K.H. "How Valid Are They? An Examination of XBRL Voluntary Filing Dcoments with the SEC EDGAR System," in: *The 14th International XBRL Conference*, Philadelphia, USA, 2006.
6. Debreceny, R.S., Chandra, A., Cheh, J.J., Guithues-Amrhein, D., Hannon, N.J., Hutchison, P.D., Janvrin, D., Jones, R.A., Lamberton, B., Lymer, A., Mascha, M., Nehmer, R., Roohani, S., Srivastava, R.P., Trabelsi, S., Tribunella, T., Trites, G., and Vasarhelyi, M.A. "Financial Reporting in XBRL on the SEC's EDGAR System: A Critique and Evaluation," *Journal of Information Systems* (29:2) 2005, pp 191-210.
7. Lee, Y.W., Strong, D., Kahn, B., and Wang, R.Y. "AIMQ: A Methodology for Information Quality Assessment," *Information & Management* (40:2) 2002a, pp 133-146.
8. Lee, Y.W., Strong., D.M., Kahn, B.K., and Wang, R.Y. "AIMQ: a methodology for information quality assessment," *Information and Management* (30:2) 2002b, pp 133-146.
9. Markus, M.L., Steinfield, C.W., Wigand, R.T., and Minton, G. "Industry-Wide Information Systems Standardization as Collective Action: The Case of the U.S. Residential Mortage Industry," *MIS Quarterly* (30:Special Issue) 2006, pp 439-465.
10. Rahm, E., and Bernstein, P.A. "A Survey of Approaches to Automatic Schema Matching," *VLDB Journal* (10:4) 2001, pp 334-350.
11. Rahm, E., Do, H.-H., and Maßmann, S. "Matching Large XML Schemas," *ACM SIGMOD Record* (33:4) 2004, pp 26-31.
12. Redman, T.C. *Data Quality for the Information Age* Artech House, Boston, 1996.
13. Rosenthal, A., Seligman, L., and Renner, S. "From Semantic Integration to Semantics Management: Case Studies and a Way Forward," *ACM SIGMOD Record* (33:4) 2004, pp 44-50.
14. van Rijsbergen, C.V. *Information Retrieval*, (2nd ed.) Butterworth, London, 1979.
15. Wang, R., and Strong, D. "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems* (12:4) 1996, pp 5-33.
16. XBRL International "Extensible Business Reporting Language (XBRL) 2.1," XBRL International, 2006.
17. Zhu, H., and Fu, L. "Towards Quality of Data Standards: Empirical Findings from XBRL," in: *The 30th International Conference on Information Systems (ICIS'09)*, Phoenix, AZ, USA, 2009.